DOCUMENT RESUME

ED 300 455 TM 012 505

AUTHOR Kingston, Neal M.; Dorans, Neil J.
TITLE Equating of GRE Aptitude Test Scores Using Item Response Theory: Effects of Multidimensionality.
PUB DATE 82
NOTE 20p.; Paper presented at the Annual Meeting of the American Psychological Association (1982).
PUB TYPE Reports - Research/Technical (143) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Aptitude Tests; *College Entrance Examinations; *Equated Scores; Higher Education; *Latent Trait Theory; *Multidimensional Scaling; *Verbal Tests
IDENTIFIERS Analytical Tests; Equipercentile Equating; *Graduate Record Examinations; Linear Equating Method

ABSTRACT

This paper described one part of an item response theory (IRT) feasibility study performed for the Aptitude Test portion of the Graduate Record Examinations (GRE). The dimensionality of the GRE verbal and analytical sections is discussed. The results of six different types of equating of the verbal section and four different types of equating of the analytical section for one form of the test, A, are compared to those of another form of the test, B. Although the absolute robustness of the various types of equating to violations of the assumption of unidimensionality cannot be determined by this research, information about the relative robustness is provided. Two forms of the GRE Aptitude Test were administered in June 1980. For Form A, the IRT analyses were based on a sample of about 3,000 examinees, while linear and equipercentile equatings were based on a sample of about 9,000 examinees. For Form B, the analyses were based on approximately 5,000 and 15,000 examinees, respectively, for IRT and non-IRT analyses. The study results are inconclusive. Simulation studies are recommended to shed light on the effects of multidimensionality on equating. Two data tables and four graphs conclude the document. (TJH)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED300455

EQUATING OF GRE APTITUDE TEST SCORES USING ITEM RESPONSE THEORY: EFFECTS OF MULTIDIMENSIONALITY

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

NEAL M. KINGSTON

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Neal M. Kingston
Selection Research
County of Los Angeles Department of Personnel

Neil J. Dorans
College Board Statistical Analysis
Educational Testing Service

A paper presented as part of the symposium, "Recent Developments in Equating at ETS", at the 1982 Annual meeting of the American Psychological Association. The programming assistance of Christopher Constantini, Edwin Kimble, and Foster Schoenthaler is gratefully acknowledged.

TM012 505

INTRODUCTION

For two or more forms of a test to be truly equated, it must be a matter of indifference for all examinees as to which form of the test they take. Lord (1980, chapter 13) shows that two test forms cannot be equated unless they are perfectly reliable or strictly parallel. The first case is not possible, and in the second case equating is not necessary. A less widely recognized requirement for equating is that the test forms to be equated be unidimensional (Morris, in press). It is, however, possible to separately equate unidimensional subsets of a test and combine the results. This would require much additional work in an operational testing environment.

Despite the impossibility of equating test forms in the strict sense described previously, one still can adjust test scores as equitably as possible. Although "as equitably as possible" needs to be defined, this paper will not do so. Instead, it will explore the potential differences in robustness to violation of the unidimensionality assumption of three equating methods, item response theory true score equating (Lord 1980, chapter 13), linear equating, and equipercentile equating (Angoff, 1971).

This paper describes one part of an item response theory feasibility study performed for the GRE Aptitude Test (for a more full description of that study see Kingston and Dorans, 1982a). In this paper we first will discuss the dimensionality of the GRE verbal and analytical sections, and then present and compare the results of six different types of equating of the verbal section and of four different typs of equating of the analytical section for one form of the test, A, to another form, B. Although the absolute robustness of the various types of equating to violations of the assumption of unidimensionality cannot be determined by this research, information about the relative robustness is provided.

DIMENSIONALITY

Studies of the dimensionality of the GRE verbal and analytical scales have been performed by Powers, Swinton, and Carlson (1977), Powers, Swinton, Thayer, and Yates (1978), Swinton and Powers (1980), Rock, Werts and Grandy (1981), Dorans and Kingston (1981), Dorans and Kingston (1982), and Kingston and Dorans (1982a). The first four studies were based on factor analysis of tetrachoric correlations. The latter two studies also used correlations of estimated item response theory parameters, and analysis of item-ability regresssions. A synopsis of selected findings of these studies follows.

Verbal Section The verbal section of the GRE aptitude test consisted of four item types: analogies, antonyms, sentence completion, and reading comprehension. One set of the reading comprehension items is based on a technically or scientifically oriented passage. The others are based on nontechnical material. Factor analytic results showed two major factors, reading comprehension and discrete verbal (all verbal item types except reading comprehension). In addition a small, probably inconsequential, technical reading comprehension factor and a reading comprehension item position factor were found. (The reading comprehension item position factor was discovered by comparing item parameter estimates from the same item given early and late in a test to equivalent groups of examinees (Dorans and Kingston 1982, Kingston and Dorans, 1982b. This analysis indicated that reading comprehension items administered at the end of a test are more difficult than those administered at the beginning of a test).

Analytical Section The analytical section of the GRE Aptitude Test consisted of three item types: analysis of explanations, logical diagrams, and analytical reasoning (for examples of the three item types see Conrad, Trismen, and Miller, 1977). Factor analysis of the binary item responses for the analytical section required six factors to explain most of the variance (Swinton and Powers, 1980). However, visual inspection of a plot of eigenvalues from the analytical item tetrachoric matrices with communality estimates in the diagonal indicated perhaps only one of these factors was psychologically meaningful (Kingston and Dorans, 1982a) and that the others may be noise factors, systematic sources of covariation among observed scores that negatively affect the fit of the factor model to the data (Tucker, Koopman, and Linn, 1969).

Analysis of item-ability regressions (Kingston and Dorans, 1982a) indicates that some of the noise factors do affect the fit of the three-parameter model to the data. Analysis of explanations item have three answer key factors. Items with a keyed response of A, C, or D form one factor; items keyed B form another, and items keyed E form a third. In addition, Kingston and Dorans (1982b) discovered a sizeable practice factor for analysis of explanations items. That is, an item appearing late in the test is much easier than the same item appearing early in the test.

EQUATING METHODS

In practice, despite efforts by test development experts, no two forms of a test can be expected to be of precisely equal difficulty. Since it is inherently unfair to compare without adjustment the raw scores of examinees who take two tests that differ in difficulty, equating procedures have been developed to transform scores from different test forms to a single scale. These equating procedures each consist of two parts, a data collection design and an analytical method to determine the appropriate transformation.

There are three basic designs for data collection: single group, equivalent-group and anchor-test (Lord, 1975). Equatings considered in this study are based on the equivalent-group design. In the equivalent-group design the old form (form already on scale) and new form (form to be scaled) are administered to random or otherwise equivalent samples from the same populations. In practice this is done through a procedure known as spiralling (Conrad, et al., 1977). Test books are packaged alternating the old and new forms and then administered within each test center so that half the examinees within each test center take each form.

Three major analytical methods to determine equating transformations were used in this research: equipercentile, linear, and item response theory based true score equating. In equipercentile equating a transformation is chosen such that scores from the two tests will be considered equated if they correspond to the same percentile rank in some group of examinees. For linear equating the chosen transformation is such that scores from the two tests will be considered equated if they correspond to the same number of standard deviations from the mean in some group of examinees. The transformation chosen for item response theory based equating is such that true scores from the two tests will be considered equated if they correspond to the same estimated theta (see Lord, 1980, chapter 13.5 for a more complete description of item response theory based true-score equating).

Although item parameter estimation and the transformation of all item parameter estimates to a single metric is a complex, integral part of item response theory equating, these issues will not be discussed in this paper. The interested reader should see Dorans (1979, 1981) or Kingston and Dorans (1982a).

DESIGN AND RESULTS

Two forms of the GRE Aptitude Test were administered operationally in June 1980. For Form A, the item response theory analyses were based upon a sample of approximately 3,000 examinees. The linear and equipercentile equatings were based on a sample of approximately 9,000 examinees. For Form B the analyses were based on approximately 5,000 and 15,000 examinees respectively for the IRT and non-IRT analyses. All sample were considered equivalent due to the spiralling of test booklets.

In addition to the operational test, each examinee took approximately half of the items from the other form in an experimental (not counted as part of the examinee's reported score) section. This experimental section was administered at the end of the test but was not identified as experimental. Thus, item parameters could be estimated based on the item's operational appearance and the item's experimental appearance at the end of the test. As stated earlier, when items are calibrated in a separate section at the end of a test they may tap an additional dimension: practice or fatigue.

Verbal Equatings

Verbal items were each calibrated twice. Once as part of the heterogenous entire verbal sections, and once just within the more homogeneous subset of either reading comprehension or discrete verbal items. These separate, more homogeneous calibrations are less likely to be affected by any violation of the unidimensionality assumption.

Form A of the verbal section was equated to Form B in six ways: linear based on equivalent groups (LE), equipercentile based on equivalent groups (EE), and four item response theory true score equatings. Three of the item response theory equatings used the assumption of group equivalence to consider the item parameter estimates to be on the same metric. The first of these used the parameter estimates based on the extant verbal sections (IES). The second used estimates based on the more homogeneous items (IESH). The third method used the heterogeneous calibrations for the old form and homogenous calibrations for the new form (IESV). The fourth item response theory true score equating was based on item calibrations from the experimental (fifth section) administration of the items (IEP).

Results of the six equatings of the verbal section form A are presented in Table 1 and Figures 1 and 2. Table 1 shows little difference in the mean scores that resulted from each of the equating variants. The exception is the IRT equating using heterogeneous calibrations for the old form and homogenous calibrations for the new form. Since this equating variant was used solely to magnify the effect of multidimensionality on equating, this is not an unexpected finding. Greater differences were manifested among the standard deviations, with the IRT equatings based on section V calibrations, heterogeneous calibrations, and homogeneous calibrations having larger standard deviations than the linear, equipercentile, and IRT heterogeneous old form, homogeneous new form calibrations. The differences among the skewnesses were fairly small. The relatively small skew of the linear equivalent groups equating might be due to the use of a linear equating model for equating two forms of different length, a situation that requires a non-linear relationship.

The comparison of means, standard deviations, and skewnesses calculated based on the different equatings is a gross analysis of the data. It should be noted that the linear and equipercentile equatings necessarily result in the same mean and standard deviation (the difference in reported standard deviations is due to rounding error) as each method sets the first two moments of the score distributions score. The one important fact Table 1 points out is multidimensionality can have a measurable effect on a single item response theory true score equating when it is magnified by using heterogeneous calibrations on one form and homogeneous calibrations on the other.

Table 1 provided a gross summary of the equating results. Figures 1 and 2 provide detail along the entire score scale. Figure 1 graphs the six equatings. Unfortunately, the resolution of the graph is insufficient to provide much detail. The only point of interest is that the linear equating is lower than the five non-linear equatings at the upper end of the score scale. As mentioned before, this is due to the difference in length between the old and new form. The linear equating model cannot fit the nonlinear relationship between the two forms.

Figure 2 magnifies the differences among equatings that could not be seen in Figure 1 by graphing the difference between each equating and the IRT equating based on items parameters estimated based on their section V appearance. Thus, the IRT section V equating (IEP) is zero across the formula score range. The graph clearly shows that for about the upper third of the formula score range (48-75) the linear equating converts to the lowest scaled score of the six equating variants. At the extreme (formula score 64 to 75) the linear equating (LE) produces a scaled score seven points lower than the next closest equating.

The differences between the IEP and IES equatings, for the most part, are small (less than one scaled score point) and are probably due to sampling error. Between formula score of 52 and 71, however, the difference ranges between two and six scaled score points. The difference between the two equatings stems solely from the position of the item within the test when it was calibrated. Dorans and Kingston have demonstrated multidimensionality within reading comprehension due to a fatigue effect (Dorans and Kingston, 1982, Kingston and Dorans, 1982b). The obtained differences are in the direction that would be expected from a fatigue effect. The reason for the localization of this effect at the upper end of the score scale is not clear but is probably idiosyncratic to the particular items in this form of the test.

The differences among the IES, IESV, and IESH equatings, for the most part, are due to the different susceptability of each method to the violation of the unidimensionality assumption. IRT models demand unidimensionality at the item level. The homogeneous calibrations come much closer to satisfying this assumption than do the heterogeneous calibrations. Thus, we expect the IESH equating to be affected least by multidimensionality, the IES equating to be affected next least, and the IESV (homogeneous new form, heterogeneous old form) to be affected most. Figure 2 supports this.

For the most part (formula scores 2 to 32 and 51 to 64) the IES equating falls between the IESH AND IESV equating. In addition, for most of the graph IESV is the most extreme IRT equating (highest equating for formula scores -8 to 38 and lowest equating for scores 51-74).

Analytical Equatings

Four different equatings of the analytical section of form A to the analytical section of form B were performed: linear equating based on equivalent groups (LE), equipercentile equating based on equivalent groups (EE), IRT true score equating based on items parametersestimated based on the item's operational appearance (IES) and IRT true score equating based on the responses to the item when it appeared in the fifth, experimental, section (IEP).

Table 2 presents the means, standard deviations, and skewnesses based on the four equatings of the analytical section. The most striking feature of Table 2 is the mean score resulting from the IRT true score equating based on the item parameters estimated from their section V appearance (IEP). The difference in means between IEP and the next closest equating is greater than one-quarter of a standard deviation (27 points). This is not surprising since this must follow from the practice effect found by Kingston and Dorans (1982b).

The other three means are quite close. The IRT true score equating with all item parameters estimated based on each item's operational administration led to a somewhat low standard deviation compared to the linear and equipercentile equatings. Why this should occur is not clear. It should be remembered, however, that the agreement between the means and standard deviations based on the linear and equipercentile equatings is required by the equating methods.

Figure 3 presents a more complete picture of the equating results. The differences among the analytical equatings are more gross than those among the verbal equatings. The one clearest fact is the IEP equating produces a considerably lower converted score than any of the other methods at every formula score above the chance level.

Figure 4, like Figure 2, magnifies the differences among the equating methods by subtracting the IEP scaled score from the scaled score obtained by each other equating method, at each formula score. This clearly shows the effect of the practice factor on IRT true score equating. Of secondary but still very important note, the other three equatings produce quite different converted scores throughout the formula score range. The difference between the linear and non-linear equatings at the upper end of the score scale is expected due to the difference in length of the two test forms. At many other parts of the score range, however, converted scores differed from 5 to 20 points.

SUMMARY

The verbal section of the GRE consists of at least two factors. Despite this, for the most part, the differences among the results of the six equating methods are small. The major exception to this is the poor results of the linear equating model at the upper end of the score scale. Since the two test forms equated differed in length, this finding was expected. Other hypotheses based on the data described in the previous section, are tentative, but other supporting (but still inconclusive) data exist (Kingston and Dorans, 1982a).

The equatings of the anlytical section produced two notable results. The practice factor is so large as to make the IEP equating variant unusable (however the makeup at this section of the GRE has been changed and this finding is likely to no longer hold). Also, once again, the linear model fairs poorly at the upper end of the score scale for test forms of unequal length.

DISCUSSION

Lord (1980, chapter 13) states that two tests cannot be equated unless they are perfectly reliable or strictly parallel. The first case is not possible and in the second case equating is not necessary. Assuming that we never have strictly parallel tests, and given the impossibility of equating fallible tests, one can still attempt to adjust scores as equitably as possible. The equating variants examined as part of this research are affected by a variety of factors. In order to judge the operational feasibility of IRT equating it is important to consider these factors and their potential though unknown effects on IRT, linear, and equipercentile equating methods.

Though Lord has shown that equating nonparallel tests requires perfect reliability, different equatings are probably differentially affected by both imperfect reliability and differences in reliability between old and new test forms. It is likely that equating methods based on true score estimates (whether they are based on IRT or classical methods) are less adversely affected, at least by differences in reliability.

Even if a lack of parallelism between test forms is attributable solely to differences in item difficulties and/or discriminations (and is unrelated to multimensionality), different equating methods will be influenced differently. Lack of statistical parallelism between form results in a curvilinear equating relationship. Since we know that we cannot produce strictly parallel tests and that, if we could, equating would be unnecessary, it is clear that linear equating can never precisely define the relationship between test scores on different forms. In many circumstances the departures from linearity appear minor, but, as test forms become less parallel, linear equating becomes less appropriate. Jaeger (1981) presents some experimental indices for investigating whether linear or equipercentile methods are more appropriate for equating.

IRT based equating, as we have chosen to implement it in this study, is not affected by speededness in the same way as is linear or equipercentile equating. To minimize multidimensionality, continguous items to which an examinee has made no response and which appear at the end of a separately timed section were coded as "not-reached" and were not used in the estimation of the examinee's ability. Likewise, these "not-reached" items were not used in estimating the parameters of the items. Thus the IRT method attempts to equate a more unidimensional ability metric. Since equating (as commonly used) provides a scaled score that is a function of an observed score, and since these observed scores have variance due to speededness, IRT equating based on item data including "not-reached" items might be subject to some problems that do not affect classical equating methods. If two forms of a test differ in speededness, IRT based equating, perhaps inappropriately, will not reflect this. The resulting bias in equating should be trivial if the variance due to the speed factor is very small compared to the variance due to the power factor or if the difference is speededness is quite small.

The equivalent-group design, as it is typically used, also presents a practical problem. When an old and new form are spiralled, the old form has previously been exposed. Some of the examinees may have previously taken the old form and thus might be expected to perform better than their fellow examinees who have either taken a different old form or have not previously taken the test. Examinees taking the new form cannot experience a comparable benefit. Thus mean scores, to some small extent, may be artificially high on the old form compared to the new form and might consistently make the old form seem easier (although probably to an unnoticeably small extent in any single equating) than it is. Such a systematic bias might lead to an eventual scale drift.

Some equating problems are based on the constraints of available data. The sparseness of data for low ability examinees makes it difficult to estimate the IRT pseudo-guessing parameter. Lack of appropriate data can also make it difficult to estimate the discrimination and difficulty parameters of very easy or very difficult items. Additionally, items that discriminate very poorly have poorly determined difficulty parameters (Kingston and Dorans, 1982a give an example of an item with parameters estimated on two samples of over 1,500 examinees where the estimate of the b parameter varied from more than +1.5 to less than -1.5). Similarly, equipercentile equating frequently suffers from a sparseness of data at the extremes of the score scale, which can lead to poor equating at those extremes. To a lesser extent, linear equating can be affected by outlying values having an undue influence on the mean and standard deviation. With the sample sizes typically used in equating the GRE Aptitude Test, however, this will not cause any difficulties.

Just as sparseness of data at the extremes presents a practical problem for some equating methods, discreteness of data can present estimation problems (Braun and Holland, in press, Potthoff, in press). Morris (in press) suggested linear equating might be preferable to equipercentile equating if there are "too few" items, but did not define "too few." Potthoff (in press) suggested not rounding formula scores before equipercentile equating or using IRT based equating to avoid problems caused by data discreteness.

The effect of multidimensionality on test equating is likely to be very complex. Investigating these effects with real data have led to mostly inconclusive results. It is recommended that simulation studies be performed to shed more light on this critical area of equating research.

REFERENCES

Angoff, W. H. Scales, norms, and equivalent scores. In. R. L. Thorndike (Ed.), Ed.) Educational measurement (2nd ed.). Washington, D. C: American Council on Education, 1971.

Braun, H. & Holland, P. Observed score test equating: a mathematical analysis of some ETS equating procedures. In P. Holland (Ed.), Proceedings of the ETS Research Statistics Conference on Test Equating. New York: Academic Press, in press.

Conrad, L., Trismen, D., & Miller, R. (Eds.) Graduate Record Examinations Technical Manual. Princeton, N. J.: Educational Testing Service, 1977.

Dorans, N. J., The need for a common metric in item bias studies. U. S. Office of Personnel Management Report TM79-20. Washington, D. C.: U. S. Office of Personnel Management, 1979.

Dorans, N. J., & Kingston, N. M. Assessing the local independence assumption of item response theory in GRE item types and populations. Paper presented at the annual meeting of the Psychometric Society, Chapel Hill, N. C., May 1981.

Dorans, N. J. & Kingston, N. M. Implications of item fungibility for innovative testing modes and equating methods. Paper presented at the annual meeting of the Psychometric Society, Montreal, June 1982.

Jaeger, R. M. Some exploratory indices for selection of a test equating method. Journal of Educational Measurement, 1981, 18, 23-38.

Kingston, N. M., & Dorans, N. J. The feasibility of using item response theory as a psychometric model for the GRE Aptitude Test. GRE Board Professional Report 79-12P, Princeton, N. J.: Educational Testing Service, 1982a.

Kingston, N. M. and Dorans, N. J. The effect of the position of an item within a test on item responding behavior: An analysis based on item response theory GRE Board Professional Report 79-12Pb. Princeton, N. J.: Educational Testing Service, 1982b.

Lord, F. A survey of equating methods based on item characteristic curve theory. Research Bulletin 75-13. Princeton, N. J.: Educational Testing Service, 1975.

Morris, C. On the foundations of test equating. In. P. Holland (Ed.), Proceedings of the ETS Research Statistics Conference on Test Equating. New York: Academic Press, in press.

Potthoff, R. Some issues in test equating. In P. Holland (Ed.), Proceedings of the ETS Research Statistics Conference on Test Equating. New York: Academic Press, in press.

Powers, D. E., Swinton, S. S., & Carlson, A. B. A factor analytic study of the GRE Aptitude Test. GRE Board Professional Report, GREB No. 75-11P. Princeton, N. J.: Educational Testing Service, 1977.

Powers, D. E., Swinton, S. S., Thayer, D., & Yates, A. A factor analytic study of seven experimental analytical item types. GRE Board Professional Report 77-78. Princeton, N. J.: Educational Testing Service, 1978.

Rock, D., Werts, C., & Grandy, J. Construct validity of the GRE across populations - an empirical confirmatory study. Draft report, 1980.

Swinton, S. S., and Powers, D. E. A factor analytic study of the restructured Aptitude Test. GRE Board Professional Report 77-6P. Princeton, N. J.: Educational Testing Service, 1980.

Tucker, L. R., Koopman, R. F., & Linn, R. L. Evaluation of factor analytic research procedures by means of simulated correlation matrices. Psychometrika, 1969, 34, 421-459.

Table 1

Verbal Equating of Form A to Form B
Means, Standard Deviations, and Skewnesses of Equated Scores
For Each Equating Variant

	Equating Variant	Mean	S. D	Skew
LE	Linear Equivalent Groups	473.29	123.35	.10
EE	Equipercentile Equivalent Groups	473.29	123.20	.20
IEP	IRT Section V Calibrations	473.81	125.47	.18
IES	IRT Heterogeneous Calibrations	473.27	125.14	.14
IESH	IRT Homogeneous Calibrations	473.39	126.51	.15
IESH	IRT Heterogeneous Old Form Homogeneous New Form Calibration	475.80	123.39	.13

Table 2

Analytical Equating of For A to Form B
Means, Standard Deviations, and Skewnessess of Equated Scores
For Each Equating Variant

Equating Variant		Mean	S.D.	Skew
LE	Linear Equivalent Groups	497.36	128.91	-.12
EE	Equipercentile Equivalent Groups	497.37	128.95	-.29
IEP	IRT Section V Calibrations	470.29	123.25	-.40
IES	IRT Heterogeneous Calibrations	498.12	125.44	-.20

Figure 1

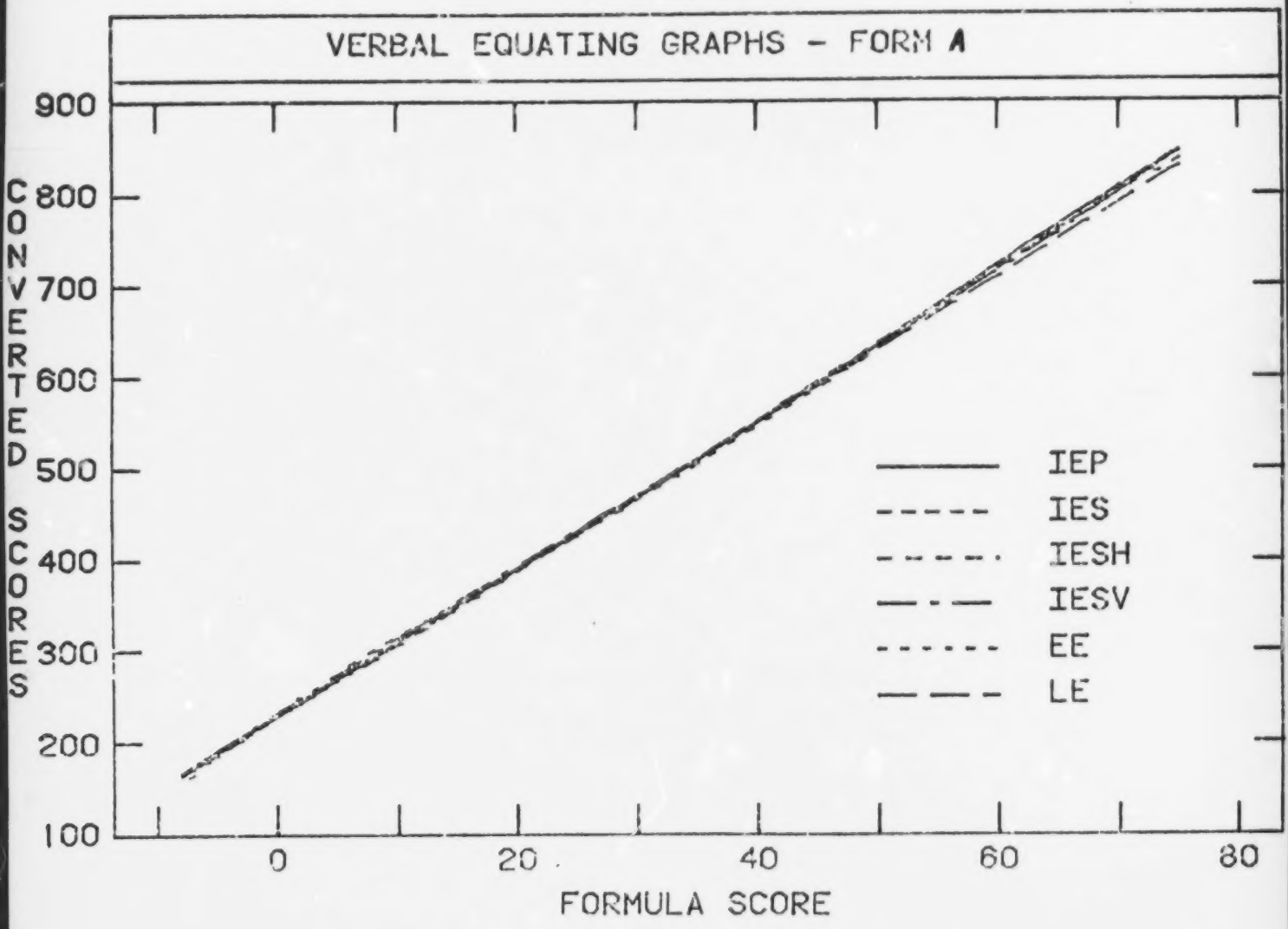
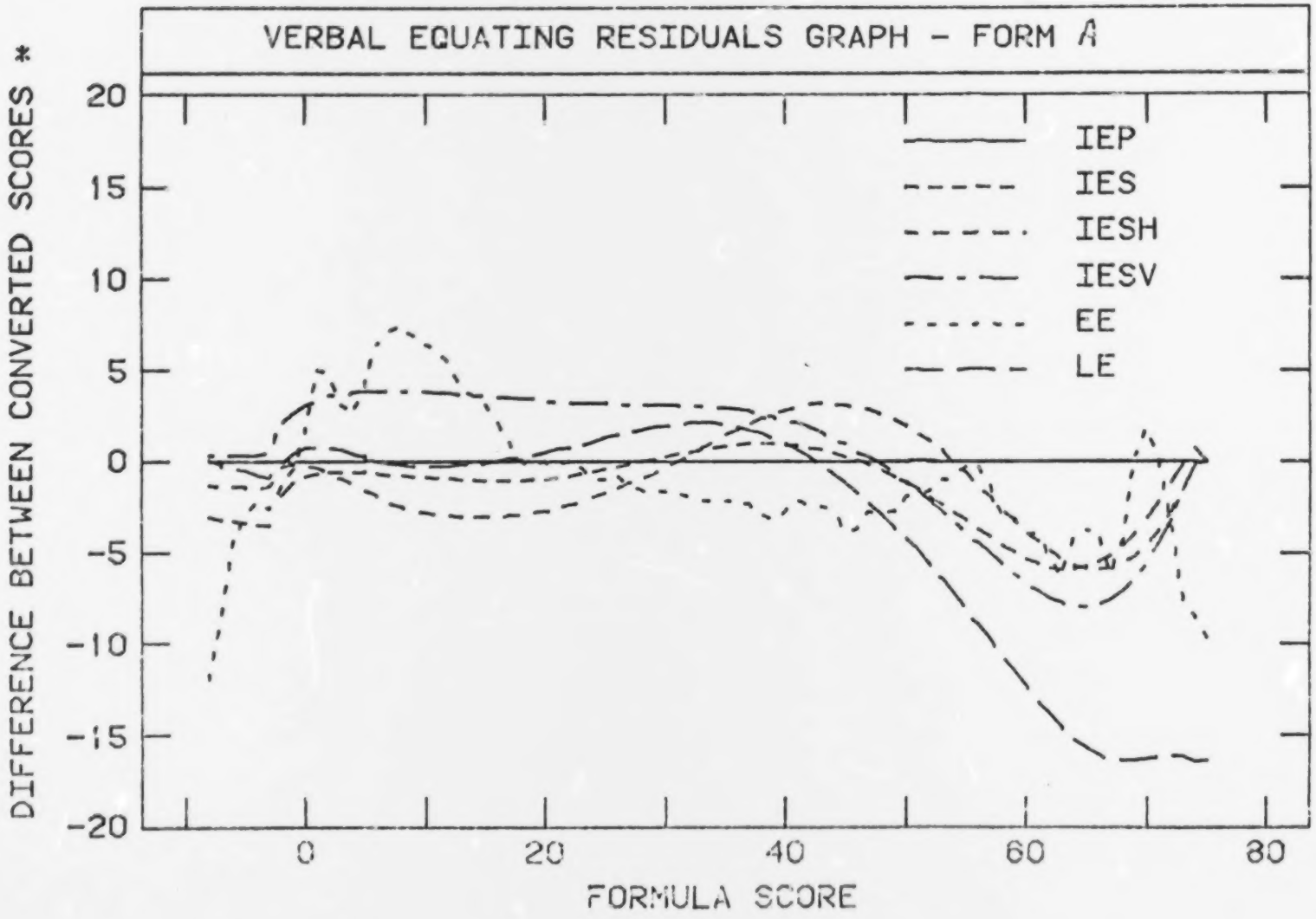

VERBAL EQUATING GRAPHS - FORM A
CONVERTED SCORES
900
800
700
600
500
400
300
200
100
0
20
40
60
80
FORMULA SCORE
IEP
IES
IESH
IESV
EE
LE

Figure 2



* SEE TEXT

Figure 3

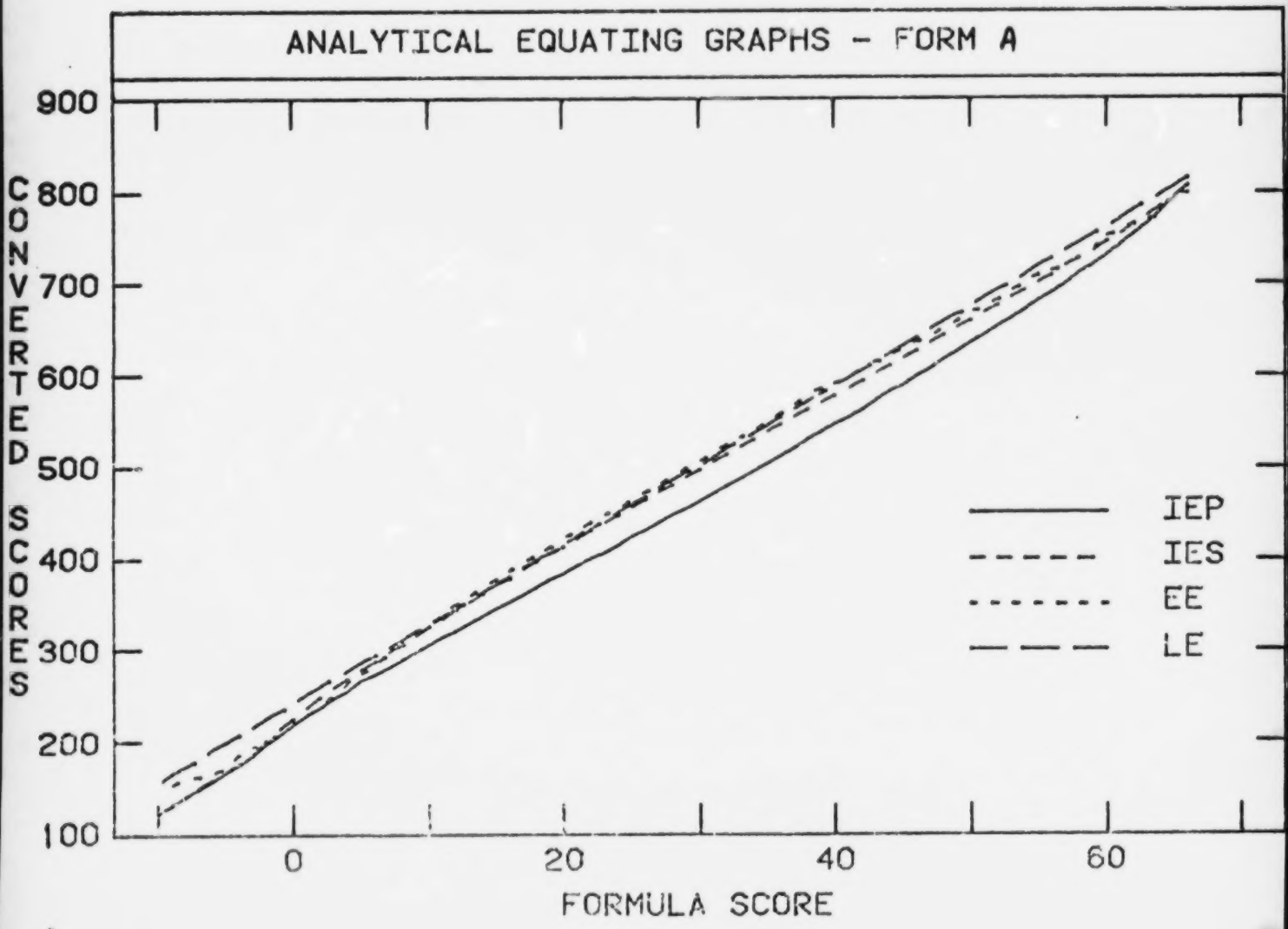
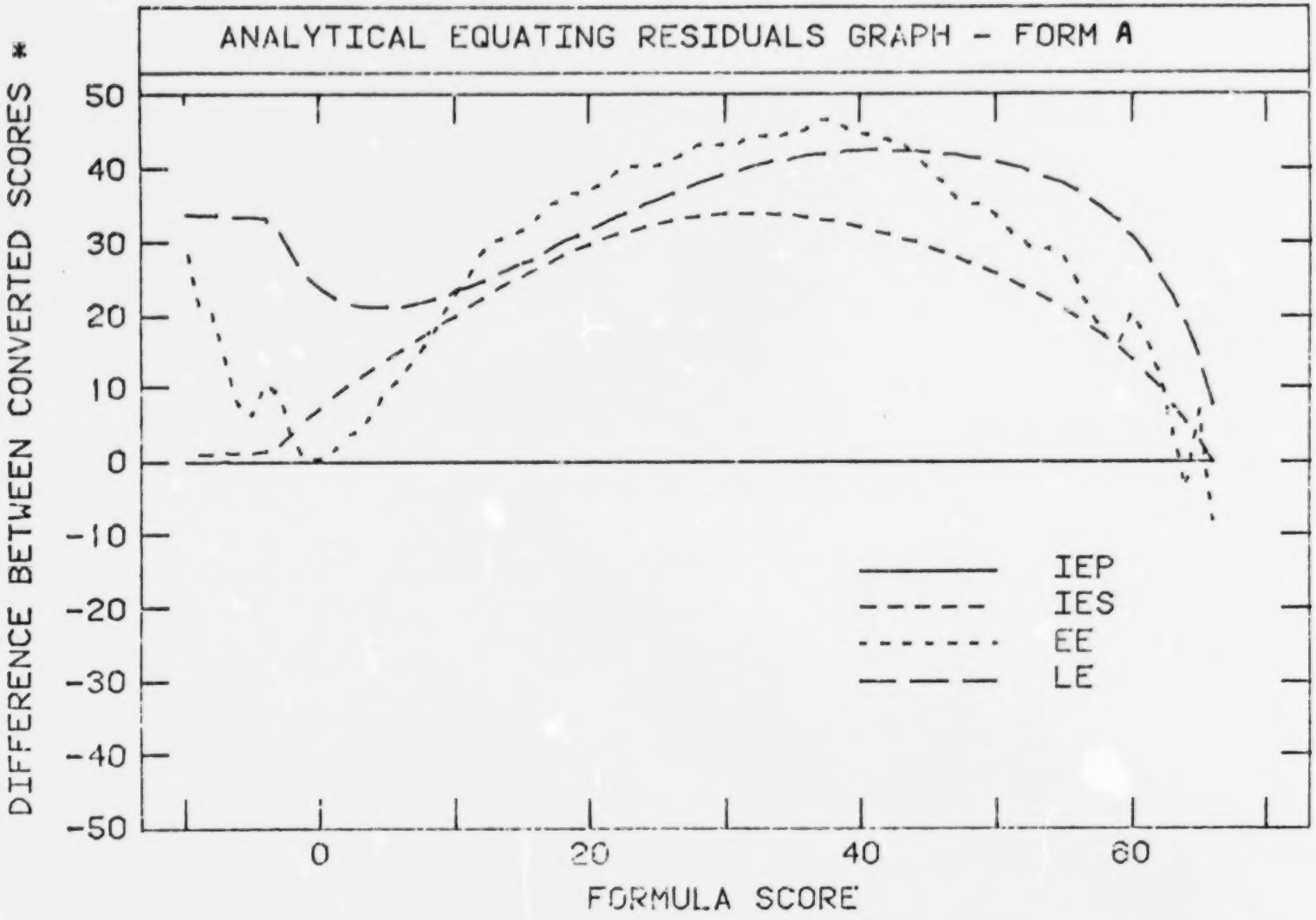

ANALYTICAL EQUATING GRAPHS - FORM A
CONVERTED SCORES
900
800
700
600
500
400
300
200
100
0
20
40
60
FORMULA SCORE
IEP
IES
EE
LE

Figure 4



* SEE TEXT